DOCUMENT RESUME

ED 138 632 TM 006 259

AUTHOR Epstein, Kenneth I.
TITLE Predictive Sample Reuse and an Application for Criterion-Referenced Item Analysis.
PUB DATE [Apr 77]
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Item Analysis; Nonparametric Statistics; Predictive Measurement
IDENTIFIERS *Predictive Sample Reuse

ABSTRACT

Predictive Sample Reuse is an alternative to parameter estimation for solving a variety of statistical problems. The technique is described and its relevance to criterion referenced measurement discussed. An item analysis procedure which does not require multiple test administrations or variability in test or item scores and which adheres to the philosophy of criterion referenced measurement more closely than do traditional correlational item analysis techniques is developed and interpreted. Directions for future theoretical and empirical research are suggested. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED138632

23.10

Predictive Sample Reuse and an Application for Criterion-Referenced Item Analysis

Kenneth I. Epstein
US Army Research Institute for the Behavioral and Social Sciences

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM006 259

Paper Presented at the Annual Meeting of the American Educational Research Association, New York, April, 1977.

Predictive Sample Reuse and an Application for Criterion-Referenced Item Analysis

Kenneth I. Epstein
US Army Research Institute for the
Behavioral and Social Sciences

The basic ideas behind this paper were suggested by several of Seymour Geisser's papers (Geisser, 1971, 1975, 1976) on the inferential use of predictive distributions. Geisser's thesis is that there has traditionally been an overemphasis in applied statistics on parameter estimation, at the expense of the development of techniques which are primarily designed to predict future observations. Predictive sample reuse (PSR) is one method for deriving and comparing mathematical predictive functions. Basically, the method requires that one or more predictive functions be evaluated on the basis of how well the values of a small sample of data from a larger data base can be predicted by the unsampled data. An item analysis technique, based on the predictive philosophy and drawing on the evaluative methodology developed by Geisser, will be described and several examples will be analyzed to illustrate the potential of the technique.

One of the most well known examples of a predictive approach to analysis is cross validation. It is common practice to validate a set of data or an analytic technique by comparing the value of some summary statistic based on one half of the data with the value based on the second half, or to correlate the first half data with the second half data. Cross validation can also be applied to the comparison of two or

The opinions expressed in this paper are those of the author and do not necessarily imply endorsement by the US Army.

more predictive functions. For example, if several point estimators were to be compared, one could compute the predicted values of each datum in one half of the data set based on the data in the second half. In this way, several predictive functions could be compared directly on their efficacy as predictors.

PSR extends the principle of cross validation in two major ways. First, it allows for partitioning of the data into as many subsets as possible rather than restricting the partition into two equal parts. The only constraint on the partitioning is that prediction to any subset of data be possible based on the remaining unsampled data. Second, PSR is a nonparametric method which may be used to analytically derive predictive functions that are optimal in some well defined predictive way.

The predictive philosophy discussed by Geisser seems to be appropriate for criterion-referenced testing. Criterion-referenced testing derives most of its benefit from its attempt to anchor test items to well defined objectives. A goal, perhaps not often realized, is that the objectives and their corresponding test items have direct relevance to the real world needs of the learner. The purpose of a criterion-referenced test is to determine an individual's level of competence relative to some objective. It seems reasonable that one operational definition of competence could involve the likelihood that an individual could successfully handle a new test item based on the same objective. Such a definition immediately suggests a predictive approach to measurement.

Two problems arise in the empirical evaluation of criterion-referenced test items. The first involves threats to the concept and purpose of criterion-referenced measurement. Millman (1974) points out that non-judicious use of item analysis results can lead to a test that does not represent the behavioral domain of interest. However, he goes on to point out that item statistics can help to detect flawed items; for example, those that are ambiguously worded, misuse graphics, or incorrectly assume student abilities beyond the scope of the domain. The second problem involves limitations of the item statistics commonly used. Indices such as the point-biserial correlation do not operate under conditions of restricted score variance and can lead to uninformative or misleading results. Other indices rely on assumptions concerning instructional effectiveness or differential ability groups and often require several test administrations. The technique proposed in this paper addresses the second problem. It requires only a single administration and does not require variability in test scores. Its predictive nature is also more within the criterion-referenced philosophy than correlational methods. However, Millman's warning against the indiscriminate use of item statistics must, of course, be heeded.

In order to use a PSR type of analysis, the data or datum to be predicted must be determined, in other words a partitioning scheme must be developed, and a measure of discrepancy between the predicted and actual values of the data or datum must be defined. The second of these problems is relatively straightforward and will be considered as part of the derivation of an item analysis index. Deciding what to predict is a more difficult conceptual problem involving some basic unresolved problems in criterion-referenced testing.

Roudabush (1974) discussed the potential disagreement between criterion-referenced measures and the definition of ability. He pointed out that both measures and abilities may be defined as either dichotomous ot multichotomous variables. The problem is that most criterion-referenced tests use dichotomously scored items regardless of whether ability is defined dichotomously or as a continuous variable. If the items measure truly dichotomous ability where masters of the objective always get the items right and nonmasters always get them wrong, then there is no conflict between the dichotomous measure and the ability being measured. If, however, ability is better described as a continuous variable, where masters get items right most of the time, say 80% or more of the time for example, and nonmasters miss items more frequently on the average, then there is a conflict between the dichotomous measure and the ability being measured. In this later case, a continuous ability is being measured by a dichotomously scored instrument. If a dichotomously scored test is used to measure dichotomously defined ability there is no basic prediction problem. There is a one to one correspondence between mastery of the objective and the item score. For the case of a dichotomously scored test used to measure continuous ability, prediction presents a problem. No item score, correct or incorrect, is a perfect predictor of the proportion of items a master or nonmaster would get correct or incorrect over repeated measurements. Each item taken individually has some built-in error associated with it. In addition, the precision of a continuous ability estimate is a function of test length. It follows, that errors in prediction will also vary with test length.

A discussion of whether ability should be defined dichotomously or continuously is beyond the scope of this paper. It is important, however, to keep in mind the different implications for interpreting item scores suggested by the two definitions of ability. An incorrect (correct) response within a string of correct (incorrect) responses suggests a poor item if one accepts the dichotomous definition of ability. That same incorrect (correct) response may or may not be a poor item from the point of view of the continuous ability model. If the item is one that high scorers consistently get wrong (low scorers consistently get right), then the item might be suspect. However, if the evidence indicates that the incorrect (correct) response was, in effect, a random error which is allowed by the continuous ability model, then the item may be fine. In fact, the occasional incorrect (correct) response is necessary for determining an individual's position on the ability continuum. Despite the possible differences in interpretation of a particular response to an item, both models imply that "good" items tend to be those which are consistently answered correctly by masters, or high scorers, and consistently missed by nonmasters, or low scorers. "Poor" items are those that are consistently missed by masters or consistently answered correctly by nonmasters or both. The item analysis technique suggested below is designed to distinguish between such "good" and "poor" items. The only assumption necessary beyond this definition of "good" and "poor" items is that at least two items are available for analysis.

The item analysis index is based on predicting a dichotomous score (0,1) for a particular person and a particular item based on that person's performance on the remaining items. Thus, for persons who tend to get most items correct, the predicted value of the item of interest is high. Conversely, for those who tend to miss most items, the predicted value will be low. The difference between the actual value of the response and the predicted value is defined as the discrepancy. The more consistent an item is with all other items, the lower the value of the discrepancy. The absolute value of the discrepancy can vary from -1 to +1. A value of -1 for the discrepancy will occur for the case of the one correct response in a series of incorrect responses. A discrepancy of +1 will occur for the case of the one incorrect response in a series of correct responses. A discrepancy of 0 will occur for all items when the responses to all of the items are either all correct or all incorrect. Notice that the discrepancy of an item is defined and calculated relative to each person. Thus, if one person gets all of the items on a test correct and a second person gets all of the items incorrect, then the item discrepancies for all items for each of these people are 0. That is, the items are working perfectly for each person even though one person gets all of the items correct while the other gets them all incorrect.

Since items may have different discrepancy values for different persons depending on their total test scores, it is desirable to calculate an average value of the discrepancy over all people for each item. Using the absolute value of the discrepancy is not suitable for such an averaging process because positive and negative values cancel each other

leading to spuriously low average discrepancy values. In order to avoid the problem, the discrepancy for each item and person is squared and the average squared discrepancy is found. The average squared discrepancy for item j will be symbolized $\bar{D}_j^2$. It can take on values from 0, for items that are completely consistent with one another, to +1 for the one inconsistent response in an otherwise consistent string of responses.

A computational formula for D is derived below. The following notation will be used:

$\bar{D}_j^2$ = the average squared discrepancy for item j

$P_{n(i\neq j)}$ = the average score for person n (proportion correct) based on all items except item j

x_{nj} = the score for person n on item j (0,1)

x_{ni} = the score for person n on item i (0,1)

I = the total number of items on the test

N = the total number of persons taking the test.

By definition,

$$\bar{D}_j^2 = 1/N \sum_{n=1}^{N} (P_{(i\neq j)} - x_{nj})^2 .$$

The computational formula is derived through some relatively simple algebra:

$$\bar{D}_j^2 = 1/N \sum_{n=1}^{N} (P_{n(i \neq j)} - x_{nj})^2$$

$$= 1/N \sum_{n=1}^{N} \left(\frac{\sum_{i=1}^{I} x_{n(i \neq j)}}{I-1} - x_{nj} \right)^2$$

$$= 1/N \sum_{n=1}^{N} \left(\frac{\sum_{i=1}^{I} x_{ni} - x_{nj}}{I-1} - x_{nj} \right)^2$$

$$= 1/N(I-1)^2 \sum_{n=1}^{N} \left(\sum_{i=1}^{I} x_{ni} - x_{nj} - x_{nj}(I-1) \right)^2$$

$$= 1/N(I-1)^2 \sum_{n=1}^{N} \left(\sum_{i=1}^{I} x_{ni} - x_{nj} - I x_{nj} + x_{nj} \right)^2$$

$$= 1/N(I-1)^2 \sum_{n=1}^{N} \left(\sum_{i=1}^{I} x_{ni} - I x_{nj} \right)^2$$

$$= 1/N(I-1)^2 \sum_{n=1}^{N} \left[\left(\sum_{i=1}^{I} x_{ni} \right)^2 - 2I x_{nj} \sum_{i=1}^{I} x_{ni} + I^2 x_{nj}^2 \right]$$

$$= 1/N(I-1)^2 \left[\sum_{n=1}^{N} \left(\sum_{i=1}^{I} x_{ni} \right)^2 - 2I \sum_{n=1}^{N} \left(x_{nj} \sum_{i=1}^{I} x_{ni} \right) + I^2 \sum_{n=1}^{N} x_{nj}^2 \right]$$

or for dichotomously scored items where $x_{nj}^2 = x_{nj}$

$$\bar{D}_j^2 = 1/N(I-1)^2 \left[\sum_{n=1}^{N} \left(\sum_{i=1}^{I} x_{ni} \right)^2 - 2I \sum_{n=1}^{N} \left(x_{nj} \sum_{i=1}^{I} x_{ni} \right) + I^2 \sum_{n=1}^{N} x_{nj} \right]$$

Note that $\sum_{n=1}^{N} \left(\sum_{i=1}^{I} x_{ni} \right)^2$ = the sum of each person's total score squared

$\sum_{n=1}^{N} \left(x_{nj} \sum_{i=1}^{I} x_{ni} \right)$ = the sum of the cross products of each person's total score and his score on item j

and $\sum_{n=1}^{N} x_{nj}$ = the sum of the responses to item j or simply the number of persons who got item j correct.

The computational formula appears a little bit formidable, particularly because of the cross product term, but it is no more difficult to apply than the formula for a point biserial correlation.

The characteristics of $\overline{D}_j^2$ can best be discussed in terms of example data. Several idealized examples will be addressed first, primarily to demonstrate some of the desirable aspects of the index. An example based on actual data will then be discussed to point out some of the limitations of $\overline{D}_j^2$, and to suggest the next steps in its development. All of the examples refer to a 12 item test administered to 30 people.

The simplest example refers to "perfect" items. Perfect item sets occur when each person's response is either all 0's or all 1's. In such cases, $\overline{D}_j^2$ will equal zero for all items regardless of the mix of completely correct and completely incorrect response patterns. The reason is obvious from the definition of the index $(1/N \sum_{n=1}^{N} (P_{n(i\neq j)} - x_{nj})^2)$. When the response pattern is all 0's, $P_{n(i\neq j)}$ will equal zero and $P_{n(i\neq j)} - x_{nj}$ will also equal zero. When the response pattern is all 1's, $P_{n(i\neq j)}$ will equal 1 and again the discrepancy will equal zero. Since the discrepancy is calculated on a per person basis and then squared and averaged, the proportion of masters and nonmasters makes no difference to the value of the index. Notice that this is in sharp contrast to the point biserial correlation which varies from zero, when the examinee group is either masters or all nonmasters, to one, when the examinee group is equally divided into masters and nonmasters.

Table 1 includes examples and comments for several more special cases. Case 1 illustrates the case of one bad item. It is written in terms of a group of masters, each of whom misses the same item. The

same results would be obtained if the group were completely composed of nonmasters, each getting the same item correct, or a mixed group as long as one and only one item was inconsistent and the item was inconsistent for all persons. For all of the "good" items, $\bar{D}_j^2$ equals .008. The value is not zero because there is a small amount of discrepancy introduced by the one "bad" item. The value of $\bar{D}_j^2$ for the one bad item is 1.0 indicating its complete lack of consistency with the other items. In this case, as in the previous case, the proportion of individuals with scores of 11 and 1 makes no difference as far as the value of $\bar{D}_j^2$ is concerned while it has a marked effect on the value of the point biserial that would be obtained.

The discussion will continue to be in terms of a group of masters. There is no loss in generality in taking this approach since the value of $\bar{D}_j^2$ is symmetric with respect to whether an example is considered from the point of view of masters or nonmasters. That is, the value of $\bar{D}_j^2$ stays the same if every 1 in the response matrix is replaced with a 0.

Case 2 (Table 1) addresses the case of all persons achieving the same total score (in this case 11 correct), with the one miss for each person randomly distributed throughout the response matrix. For the case of 30 persons each taking a 12 item test this means that 6 items (say 1 through 6) are missed by two persons each and 6 items (say 7 through 12) are missed by three persons each. In the first case $\bar{D}^2_{j=1-6}$ equals .074, in the second case it equals .107. A comparison of this case with case 1 reveals that none of the case 2 items are as "good" as the best items in case 1. Nor do their index values approach the value of 1.0 for the "bad" item in case 1. Rather, all of the items in case 2 show a small discrepancy. The difference in the index values

for items 1 through 6 and items 7 through 12 is attributable to the number of persons for whom inconsistencies arise. Cases 3 through 6 in Table 1 are similar to cases 1 and 2, differing only in the number of items missed. As expected, the value of $\bar{D}_j^2$ increases as the number of inconsistencies increases, with smaller values within a total score group when specific "bad" items can be identified (cases 3 and 5), and slightly larger values when misses are distributed across all of the items.

For the case of homogeneous score groups, the value of $\bar{D}_j^2$ is predictable and relatively easy to interpret. For example, consider case 6 on Table 1. In case 6, each person's score is 9, with the 90 (3 x 30) 0's in the response matrix randomly distributed across all items. This implies that 6 items (say items 1 through 6) are missed by seven persons each (= 42 misses), and 6 items (say items 7 through 12) are missed by eight persons each (= 48 misses). The calculation of $\bar{D}_j^2$ is straightforward:

$$\bar{D}_j^2 = 1/N(I-1)^2 \left[\sum_{n=1}^{N} \left(\sum_{i=1}^{I} x_{ni} \right)^2 - 2I \sum_{n=1}^{N} \left(x_{nj} \sum_{i=1}^{I} x_{ni} \right) + I^2 \sum_{n=1}^{N} x_{nj} \right]$$

where $1/N(I-1)^2 = 1/30(11)^2 = 1/3630$

$\left(\sum_{i=1}^{I} x_{ni} \right)^2 = 9^2 = 81$ for all persons

$\sum_{n=1}^{N} \left(x_{nj} \sum_{i=1}^{I} x_{ni} \right) = \begin{cases} 23(9) \text{ for items } 1\text{-}6 \\ 22(9) \text{ for items } 7\text{-}12 \end{cases}$

$\sum_{n=1}^{N} x_{nj} = \begin{cases} 23(1) \text{ for items } 1\text{-}6 \\ 22(1) \text{ for items } 7\text{-}12 \end{cases}$

$$\bar{D}^2_{j=1\text{-}6} = 1/3630 \left[30(81) - 24(23(9)) + 144(23) \right] = \frac{774}{3630} = .213$$

$$\bar{D}^2_{j=7\text{-}12} = 1/3630 \left[30(81) - 24(22(9)) + 144(22) \right] = \frac{846}{3630} = .233$$

Notice the final results of these calculations. The denominator in each case is simply a constant averaging factor. The value of $\bar{D}_j^2$ is actually determined by the value of the numerator. For homogeneous groups the numerator has a direct interpretation in terms of the number of persons missing an item.

Let x equal the number of persons who miss an item, and y equal the number who get it right. The squared discrepancies for incorrect and correct responses are, by definition, $(9/11 - 0)^2 = (9/11)^2$ and $(8/11 - 1)^2 = (-3/11)^2$ respectively. Hence, $\bar{D}_j^2 = [x(81/121) + y(9/121)]/30 = (81x + 9y)/3630$. The numerator can be equated to the number of persons getting the item incorrect and correct. Since the total number of persons is known, two equations are available which can be solved simultaneously:

For items 1 through 6:

$$81x + 9y = 774,$$
$$x + y = 30,$$

and $x = 7$, $y = 23$

For items 7 through 12:

$$81x + 9y = 846,$$
$$x + y = 30,$$

and $x = 8$, $y = 22$.

These values correspond to the initial values used to define this example.

Unfortunately, for heterogeneous examinee groups the analysis is not so straightforward. The reason is that two unknowns are associated with each score group, the number who get an item right and the number who get it wrong, and a unique solution to the resulting set of linear equations does not appear possible. Some initial steps in defining a scale of $\bar{D}_j^2$ values have been made. As a first step, the per person squared discrepancies for all score groups given a 12 item test were calculated. The results are shown on Table 2. Two other sets of

calculations, Tables 3 and 4 and Figures 1 and 2, show values of the squared discrepancy as a function of the identity and number of score groups for which an item seems to be working correctly. These calculations are all based on the 12 item test. While the same relationships will exist between $\bar{D}_j^2$ values for similar conditions assuming different length tests, the actual values will vary. Sample size also affects the value of the index for a specific example. The example calculations also assume that when the examinee group is characterized by heterogeneous score groups that the proportion of examinees in each group is equal. The effect on the value of the index when the equal proportionality assumption is not true has yet to be systematically studied. The purpose of the above examples is to illustrate the potential of the squared discrepancy index. Clearly, there is much to be done before it can be generally used.

Finally, some real, if not very cooperative, data are considered. The data are the responses to an entry placement test in basic mathematics used at an adult basic education center in Washington, D.C. The data are found in Table 5. Values for $\bar{D}_j^2$, r_{jT}, and the rank order of the items on each of these indices is shown on Table 6. The 12 items are all free response and include problems of addition, subtraction, multiplication, and division of integers, fractions, and decimals, and working with percentages. The test is used to determine which students can begin classes aimed at preparation for the GED examination and which students must enter a remedial basic skills class.

Relative to the hypothesis that $r_{jT} = 0$, the probabilities for the values obtained for items 5,6,12, and 4 are less than .01, for items 3 and 10, less than .05, and for the other items the probability is greater than .05. In much the same way that an idealized distribution is referred to when interpreting values of r_{jT}, the idealized values for $\overline{D}_j^2$ represented by Tables 3 and 4 provide some insight in interpreting the obtained values.

Table 3 shows the values of the squared discrepancy as items begin to work for extreme score groups. The greatest discrepancies are found for "bad" items in extreme score groups with smaller discrepancies associated with score groups approaching half the items correct. This is simply because at the half correct score group level a miss is equivalent to a correct response in terms of its consistency with the other items on the test, while at the extremes, misses (or correct responses) are extremely inconsistent with the other responses. For criterion-referenced testing purposes, the data will often represent extreme groups; masters or nonmasters. Hence, the greatest concern might be that items be most effective for such groups.

The smallest average squared discrepancy found in the items in the example data set was .218, the largest value was .330. Referring to Table 3, a value of .218 is equal to the value obtained for items working at the three most extreme scores in each direction. That is, for those with total scores of 1, 2, 3, 9, 10, or 11. Perfect or completely incorrect scores are not included since their discrepancies are zero. Such items are likely to be acceptable for most decision making purposes. The "worst" item had an average squared discrepancy of .330.

This value is slightly worse than the value that would be obtained if only the two most extreme score groups in each direction were being adequately handled. Such items may be bordering on the edge of usefulness for decision making purposes.

Table 4 approaches the scaling problem from the central score group moving out. That is, it shows the values of the average squared discrepancy as items begin working for score groups in the center, with maximum discrepancy remaining with respect to extreme scores. With respect to Table 4, the best items from the example are approximately equal to items that work for all but the most extreme score groups, 1 and 11. The worst item is still operating well for central scores. It approximates the value that would be obtained if the seven central score groups and either score group 2 or group 10 had a minimal number of inconsistencies.

The purpose of this paper was to introduce a predictive approach to analysis and to describe a first attempt to apply that approach to item analysis. While it seems premature to suggest that the particular item analysis technique outlined can be generally applied, the technique does present interesting and potentially useful possibilities. There is no question that the average squared discrepancy is a meaningful index for certain special cases. That those cases include instances of extremely restricted score variance and widely separated bimodal distributions is encouraging since applications of criterion-referenced tests often lead to such results. The most critical issue to be resolved is the interpretation of $\overline{D}_j^2$ values in terms of the effectiveness of items for more general cases.

References

Geisser, S. The inferential use of predictive distributions. In Godambe, V. P. and Sprott, D. A. (Eds.), Foundations of statistical inference, Toronto, Montreal: Holt, Rinehart, and Winston. 1971.

Geisser, S. The predictive sample reuse method with applications. J. Amer. Statis. Asso. 1975, 70, 350, 320-328.

Geisser, S. Predictivism and sample reuse. Proceedings of the Twenty-First Conference on the Design of Experiments, Research Triangle Park, North Carolina: US Army Research Office, 1976, 349-362.

Millman, J. Criterion-referenced measurement. In Popham, W.J. (Ed.) Evaluation in education: Current applications. Berkeley, Cal.: McCutchen Publishing Corporation, 1974.

Roudabush, G. E. Models for a beginning theory of criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, 1974

CASE	DESCRIPTION	$\bar{D}^2_j$	r_{iT}
1	All persons get items 1 through 11 correct; all persons miss item 12		
	(a) items 1 through 11	30/3630 = .008	0.0
	(b) item 12	3630/3630 = 1.0	0.0
2	All persons get 11 items correct; the incorrect responses are random		
	(a) items missed by two persons each (6 items)	270/3630 = .074	undefined
	(b) items missed by three persons each (6 items)	390/3630 = .107	undefined
3	All persons get items 1 through 10 correct; all persons miss items 11 and 12		
	(a) items 1 through 10	120/3630 = .033	0.0
	(b) items 11 and 12	3000/3630 = .826	0.0
4	All persons get 10 items correct; the incorrect responses are random (each item is missed by five persons)	600/3630 = .167	undefined
5	A persons get items 1 through 9 correct; all persons miss items 10 through 12		
	(a) items 1 through 9	270/3630 = .074	0.0
	(b) items 10 through 12	2430/3630 = .669	0.0
6	All persons get 9 items correct: the incorrect responses are random		
	(a) items missed by seven persons each (6 items)	774/3630 = .213	undefined
	(b) items missed by eight persons each (6 items)	846/3630 = .233	undefined
7	Persons 1 and 2 miss all items: person 3 misses 6 items (items 7 through 12): persons 4 through 30 get all items correct		
	(a) items 1 through 6	36/3630 = .010	.939
	(b) items 7 through 12	36/3630 = .010	.958
8	Persons 1 through 5 miss items 4 through 12; persons 6 through 10 miss items 10 through 12; persons 11 through 30 get all items correct		
	(a) items 1 through 3	450/363) = .124	0.0
	(b) items 4 through 9	90/3630 = .025	.944
	(c) items 10 through 12	450/3630 = .124	.853

Table 1: Values of $\bar{D}^2_j$ for several example "special" cases assuming a 12 item test administered to 30 persons

Total Number Correct	Incorrect Response	Correct Response
12	-----	-----
11	$(11/11)^2 = 121/121 = 1.00$	$(-1/11)^2 = 1/121 = .008$
10	$(10/11)^2 = 100/121 = .826$	$(-2/11)^2 = 4/121 = .033$
9	$(9/11)^2 = 81/121 = .669$	$(-3/11)^2 = 9/121 = .074$
8	$(8/11)^2 = 64/121 = .529$	$(-4/11)^2 = 16/121 = .132$
7	$(7/11)^2 = 49/121 = .405$	$(-5/11)^2 = 25/121 = .207$
6	$(6/11)^2 = 36/121 = .298$	$(-6/11)^2 = 36/121 = .298$
5	$(5/11)^2 = 25/121 = .207$	$(-7/11)^2 = 49/121 = .405$
4	$(4/11)^2 = 16/121 = .132$	$(-8/11)^2 = 64/121 = .529$
3	$(3/11)^2 = 9/121 = .074$	$(-9/11)^2 = 81/121 = .669$
2	$(2/11)^2 = 4/121 = .033$	$(-10/11)^2 = 100/121 = .826$
1	$(1/11)^2 = 1/121 = .008$	$(-11/11)^2 = 121/121 = 1.00$
0	-----	-----

Table 2: Per Person Squared Discrepancies for Incorrect and Correct Responses as a Function of Total Score for a 12 Item Test

Case	Score Groups Included	$\overline{D}_j^2$
a	none	.651
b	11 OR 1	.560
c	11 AND 1	.470
d	1, 10, 11 OR 1, 2, 11	.398
e	1, 2, 10, 11	.326
f	1, 2, 9, 10, 11 OR 1, 2, 3, 10, 11	.272
g	1, 2, 3, 9, 10, 11	.218
h	1, 2, 3, 8, 9, 10, 11 OR 1, 2, 3, 4, 9, 10, 11	.182
i	1, 2, 3, 4, 8, 9, 10, 11	.146
j	1, 2, 3, 4, 7, 8, 9, 10, 11 OR 1, 2, 3, 4, 5, 8, 9, 10, 11	.128
k	ALL	.110

Table 3: $\overline{D}_j^2$ Values for a 12 Item Test as a Function of Score Groups for Which Item j Squared Discrepancies are Minimal

(In all cases, squared discrepancies for non-included groups are maximal; equal proportions of persons across score groups is assumed; score groups 0 and 12 are excluded since they do not contribute to the squared discrepancy value)

Case	Score Groups Included	$\bar{D}_j^2$
a	6	.651
b	6,7 OR 5,6	.633
c	5, 6, 7	.615
d	4, 5, 6, 7 OR 5, 6, 7, 8	.578
e	4, 5, 6, 7, 8	.542
f	3, 4, 5, 6, 7, 8 OR 4, 5, 6, 7, 8, 9	.488
g	3, 4, 5, 6, 7, 8, 9	.434
h	2, 3, 4, 5, 6, 7, 8, 9 OR 3, 4, 5, 6, 7, 8, 9, 10	.362
i	2, 3, 4, 5, 6, 7, 8, 9, 10	.290
j	1, 2, 3, 4, 5, 6, 7, 8, 9, 10 OR 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	.200
k	ALL	.110

Table 4: $\bar{D}_j^2$ Values for a 12 Item Test as a Function of Score Groups for which Item j Squared Discrepancies are Minimal

(In all cases, squared discrepancies for non-included groups are maximal; equal proportions of persons across groups is assumed; score groups 0 and 12 are excluded since they do not contribute to the squared discrepancy value)

	ITEM												
PERSON	1	2	3	4	5	6	7	8	9	10	11	12	TOTAL
1	1	1	0	0	1	1	0	1	1	0	1	1	8
2	1	1	1	0	0	0	0	1	0	0	0	0	4
3	1	1	1	0	1	1	1	1	1	0	1	1	10
4	1	1	0	0	0	0	0	1	1	0	0	1	5
5	1	1	0	0	0	0	1	1	1	0	0	0	5
6	1	1	1	1	1	1	1	0	1	1	0	0	9
7	1	1	1	0	0	0	1	1	1	0	0	0	6
8	1	1	0	0	0	0	0	1	1	1	0	1	6
9	1	1	0	1	0	0	1	1	0	0	0	0	5
10	1	1	0	1	1	1	1	1	1	1	0	1	10
11	1	0	1	1	1	1	1	1	0	0	0	1	8
12	1	0	1	0	0	1	0	1	1	0	1	1	7
13	1	1	1	1	0	0	1	1	1	0	0	0	7
14	1	1	1	1	0	0	1	1	0	0	0	1	7
15	1	1	0	0	1	1	1	1	0	0	0	0	6
16	1	0	0	0	0	0	0	1	0	0	0	0	2
17	1	1	0	1	0	0	0	1	1	1	0	1	7
18	1	1	1	1	1	1	1	1	1	1	0	1	11
19	1	1	1	0	0	0	1	1	1	0	0	0	6
20	1	1	1	0	0	1	0	1	1	0	0	1	7
21	1	1	0	1	0	0	0	1	1	1	0	1	7
22	1	1	1	0	0	0	0	1	1	0	0	0	5
23	1	1	1	1	0	0	0	1	1	0	1	0	7
24	1	1	1	1	0	0	1	1	1	0	0	0	7
25	1	0	1	1	1	1	1	1	1	0	0	1	9
26	1	0	0	0	0	0	0	0	1	0	0	0	2
27	1	1	1	0	0	0	0	1	1	1	0	0	6
28	0	0	0	0	0	0	1	1	1	0	0	0	3
29	1	0	0	1	0	0	1	1	1	0	0	0	5
30	1	1	1	1	0	0	1	1	0	0	0	1	7
SUM	29	23	17	14	8	10	17	28	23	7	4	14	194

$\bar{Y} = 6.47$ $s_y^2 = 4.67$

Table 5: Data collected on 30 subjects for a 12 item test of basic arithmatic skills

Item	$\bar{D}_j^2$	Rank	r_{jT}	Rank
1	.271	(9)	.303	(9)
2	.251	(6)	.344	(8)
3	.244	(5)	.414	(5)
4	.237	(4)	.486	(4)
5	.231	(3)	.683	(1)
6	.218	(1)	.677	(2)
7	.257	(7)	.350	(7)
8	.284	(10)	.122	(12)
9	.271 = item 1	(8)	.232	(11)
10	.290	(11)	.398	(6)
11	.330	(12)	.283	(10)
12.	.218 = item 6	(2)	.581	(3)

Table 6: Item $\bar{D}_j^2$ and r_{jT} Values and Ranks for the Example Data

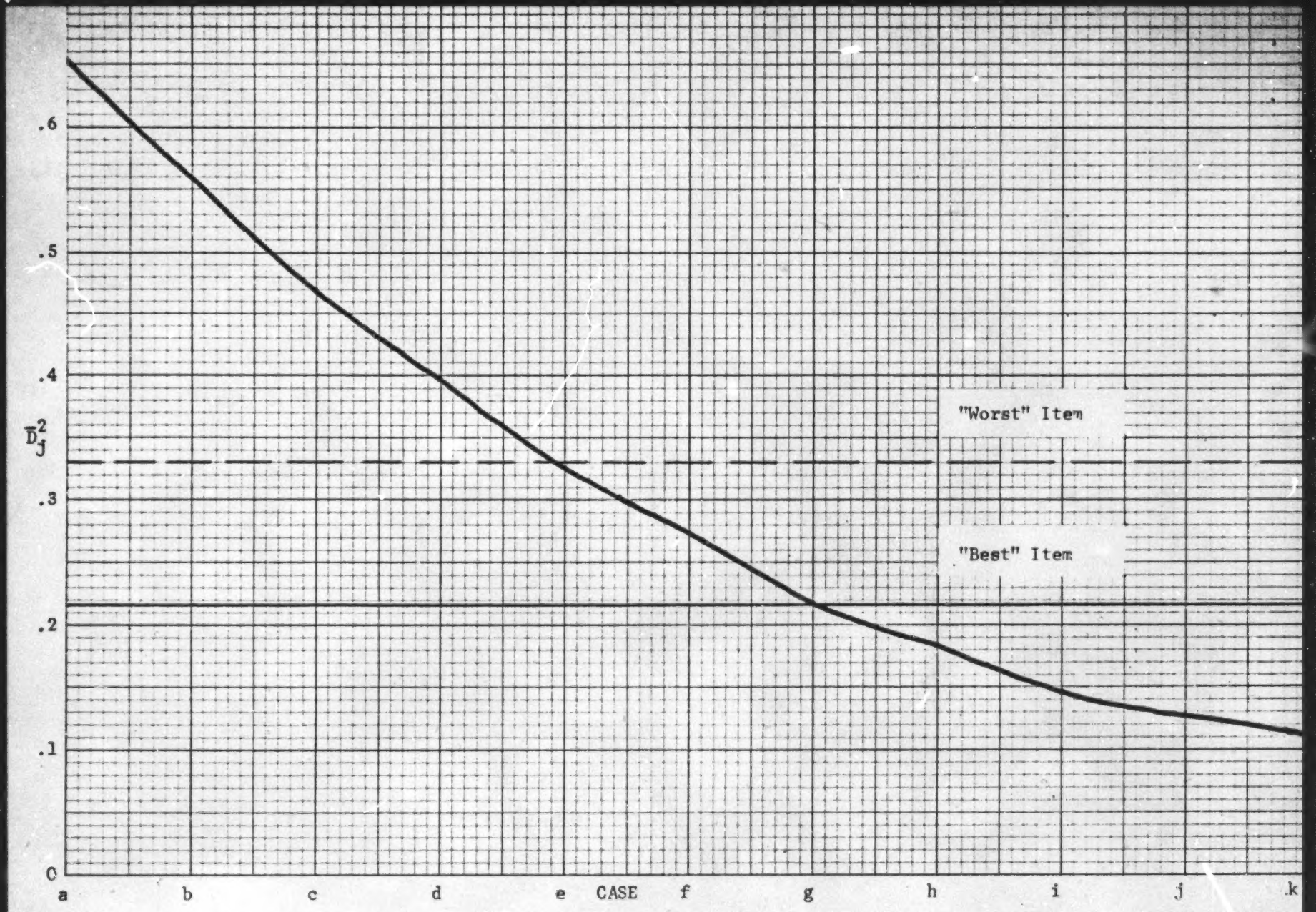
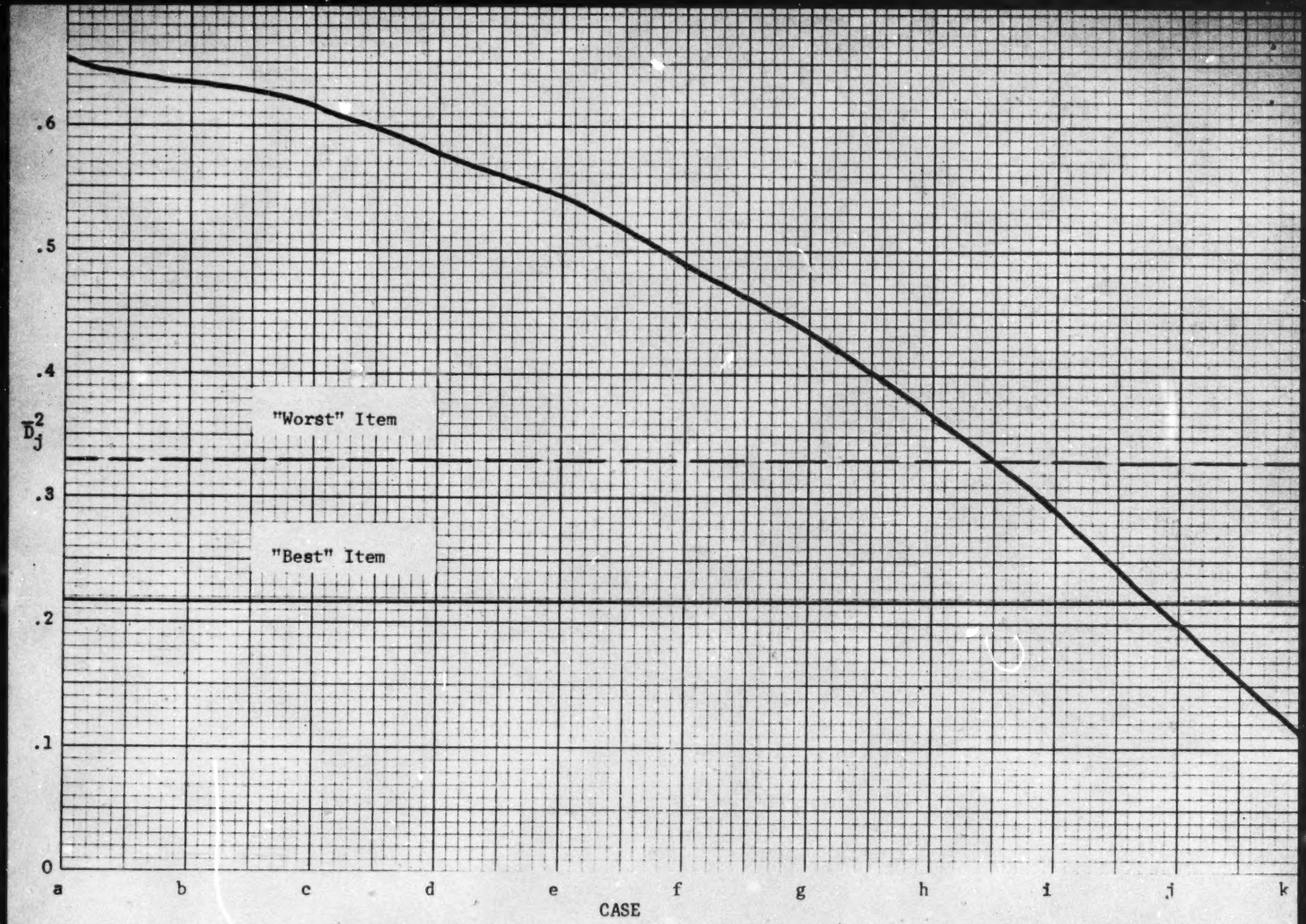


Figure 1: Table 3 Data



27 Figure 2: Table 4 Data